pK_a measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments

Mehtap Işık[1,2], Dorothy Levorse[3], Ariën S. Rustenburg[1,4], Ikenna E. Ndukwe[5], Heather Wang[6], Xiao Wang[5], Mikhail Reibarkh[5], Gary E. Martin[5], David Mobley[7], Timothy Rhodes[3], John D. Chodera[1*]

[1]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, United States; [2]Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065, United States; [3]Merck & Co., Inc., MRL, Pharmaceutical Sciences, 126 East Lincoln Avenue, Rahway, New Jersey 07065, United States; [4]Graduate Program in Physiology, Biophysics, and Systems Biology, Weill Cornell Medical College, New York, NY 10065, United States; [5]Merck & Co., Inc., MRL, NMR Structure Elucidation, 126 East Lincoln Avenue, Rahway, New Jersey 07065, United States; [6]Merck & Co., Inc., MRL, Process Research & Development, 126 East Lincoln Avenue, Rahway, New Jersey 07065, United States; [7]Department of Pharmaceutical Sciences and Department of Chemistry, University of California, Irvine, Irvine, California 92697, United States

***For correspondence:**
john.chodera@choderalab.org (JDC)

Abstract Determining the protonation state of a small molecule is a preliminary requirement for predicting its physicochemical and pharmaceutical properties, as well as interactions with protein targets using computational models. To determine the ionic state of a molecule in an aqueous solution at a certain pH it is necessary to know its acid dissociation constants(pKa values). As a part of SAMPL6 community challenge, we organized a blind pKa prediction component to asses the accuracy of contemporary pKa prediction methods. While inaccuracy in prediction of small molecule pKas have potential detrimental impact on predictive physical models, predicting pKas of drug-like molecules can be difficult due to challenging properties such as multiple titratable sites, heterocycles, and tautomerization. We limited the focus of this challenge on a subset of chemical space of drug-like molecules: 24 small molecules were selected to represent fragments of kinase inhibitors. We measured macroscopic pKa values of SAMPL6 compounds with UV-absorbance based method with Sirius T3 instrument to construct an experimental reference dataset for the evaluation of computational pKa predictions.

Keywords

acid dissociation constants · spectrophotometric pKa measurement · blind prediction challenge · SAMPL · macroscopic pKa · microscopic pKa · macroscopic protonation state · microscopic protonation state

Abbreviations

SAMPL Statistical Assessment of the Modeling of Proteins and Ligands

pKa $-log_{10}$ acid dissociation equilibrium constant

psKa $-log_{10}$ apparent acid dissociation equilibrium constant in cosolvent

DMSO Dimethyl sulfoxide

ISA Ionic-strength adjusted

SEM Standard error of the mean

TFA Target factor analysis

LC-MS Liquid chromatography - mass spectrometry

NMR Nuclear magnetic resonance spectroscopy

Introduction

SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) is a series of annual blind prediction challenges for the computational chemistry community. It aims to evaluate and advance computational tools for rational drug design. In addition to protein-ligand binding affinity predictions, evaluating methods for predicting relevant physicochemical properties like solvation free energies and distribution coefficients have been part of SAMPL challenges and provided efficient tests for different aspects of physical modeling based and empirical predicting methods, such as effect of force field accuracy, solvation models, pKa and tautomer predictions. Physicochemical property prediction challenges have been observed to be helpful to pinpoint deficiencies in computational models that can lead to substantial errors in affinity predictions.

Cyclohexane-water distribution constant (logD) prediction challenge was organized as a part of SAMPL5 in the previous year [1, 2]. Conformational sampling of flexible solute molecules, predicting relevant protonation and tautomarization states, bias towards cyclohexane phase related to force field accuracy and modeling of the solvent were some of the potential contributors to inaccuracy of logD predictions, justifying the benefit of organizing future iterations of blind distribution or partition coefficient challenges. Moreover, observations in logD prediction challenge indicated that neglecting protonation state effects leads to large errors in transfer free energies [1–3]. Protonation state effects contributed to the error up to 6-8 kcal/mol for some compounds [3]. To isolate the acid dissociation constant (pKa) prediction problem from difficulties related to lipophilicity (logP and logD) prediction methods, we decided to direct a set of staged physicochemical property challenges using the same set of molecules: Initially, a pKa prediction challenge was conducted while keeping collected experimental pKa data blind. In the future, with experimental pKa values provided blind a computational partition coefficient challenge will be conducted. This article reports on the experimental measurements for SAMPL6 pKa challenge.

This is the first time, a blind pKa prediction challenge is fielded in SAMPL. By organizing a computational pKa prediction challenge in SAMPL6, we aimed to capture the performance of current pKa prediction methods and isolate potential detriment of inaccurate pKa estimates for predicting affinities of drug-like molecules. Similar to logD predictions, any error in prediction of the protonation free energy of an ionizable ligand can directly add on to the error in binding free energy calculations for a protein-ligand complex. pKa prediction of drug-like molecules are complicated by the following molecular properties: multiple protonation sites, heterocycles, tautomerization, conformational flexibility of large molecules and intramolecular hydrogen bonds. Since the scope of experimental data collection was on the order of tens of compounds, we decided to focus on the chemical space of kinase inhibitors in the first iteration of pKa prediction challenge. 24 small organic molecules (17 fragment-like and 7 drug-like) were selected for their similarity to kinase inhibitors and considering experimental tractability of pKa and logP measurements. Macroscopic pKa values were collected experimentally with UV-absorbance spectroscopy based pKa measurements with Sirius T3. Experimental data was kept for 3 months (Oct. 25, 2017 - Jan. 23, 2018) to allow blind computational predictions. 11 research groups participated in this challenge with 93 prediction submission sets that cover a large variety of contemporary pKa prediction methods.

Whenever experimental pKa measurements are used for evaluating pKa predictions, it is important to differentiate between microscopic and macroscopic pKa values. In molecules with multiple titratable groups, the protonation state of one group can affect the proton dissociation propensity of another functional group. In such cases, the **microscopic pKa** refers to the pKa of deprotonation of a single titratable group while all the other titratable and tautomerizable functional groups of the same molecule are held fixed. Different protonation states and tautomer combinations constitute different microstates. The **macroscopic pKa** defines the acid dissociation constant related to the loss of a proton from a molecule regardless of which functional group the proton is dissociating from, so it doesn't necessarily convey structural information.

Weather a measured pKa is microscopic or macroscopic depends on the experimental method used (Figure 1). For a molecule with only one protonatable group, the microscopic pKa is always equal to the macroscopic pKa. But for a molecule with multiple titratable groups, throughout a titration from acidic to basic pH, the deprotonation of some functional groups can take place almost at the same time. Then the experimentally measured macroscopic pKa will have contributions from multiple microscopic pKas with similar values (i.e. acid dissociation of multiple microstates). Cysteine provides an example of this behavior with its two macroscopic pKas observable by spectrophotometric or potentiometric pKa measurement experiments [4]. While 4 microscopic pKas can be defined for cysteine, experimentally observed pKas can't be assigned to individual functional groups directly and requires more advanced techniques such as NMR. On the other hand, when there is a large difference between microscopic pKas of a molecule, the proton dissociations won't overlap and macroscopic pKas observed by experiments can be assigned to individual titratable groups. The pKa values of glycine provide a good example of this scenario [4].

UV-absorbance spectroscopy (UV-metric titration), potentiometry (pH-metric titration), capillary electrophoresis , and NMR spectroscopy are most common pKa measurement methods [7]. UV-metric and pH-metric methods (Figure 2) are applicable to measurement of aqueous pKa values between 2 and 12, due to limitations of pH-electrode. The pH-metric method relies on determining the stochiometry of bound protons with respect to pH, calculated from volumetric titration of acid or base solutions. Accurate measurements require high concentration of analyte and analytically prepared acid/base stocks and analyte solutions. The UV-metric pKa measurements rely on difference in UV-absorbance spectra of different protonation states. pH and UV absorbance of analyte solution is monitored during titration.

find reference

Both UV-metric and pH-metric pKa determination methods measure macroscopic pKas for polyprotic molecules, which can not be easily assigned to individual titration sites and underlying microstate populations in the absence of other experimental evidence that provides structural information, such as NMR (Figure 1). Macroscopic populations observed in these two methods are composed of different combinations of microstates depending on the principles of measurement technique. In potentiometric titrations microstates with same total charge will be observed as one macrostate, while in spectrophotometric titrations some protonation of some titration sites could be invisible and macrostates will be formed of sum of microstate populations that manifest similar UV-absorbance spectra.

Spectrophotometric pKa method is more sensitive than potentiometric method and requires very low analyte concentrations(∼ 50 µM) which is advantageous especially for compounds with low solubility, but it is only applicable to titration sites near chromophores. For protonation to effect absorbance protonation site should be a maximum of 4 heavy atoms away from the chromophore: conjugated double bonds, carbonyl groups, aromatic rings etc. Although potentiometric measurement doesn't have this structural limitation, a higher concentration of analyte (∼ 5 mM) is necessary for potentiometric method than for spectrophotometric method to provide large enough buffering capacity signal above water for an accurate measurement. Accuracy of pKa calculation from pH-metric method is also sensitive to correct estimation of analyte concentration in the sample solution. Therefore, we have decided to use spectrophotometric measurements for collecting experimental pKa data and selected SAMPL6 pKa challenge compounds so that all their potential titration sites are in the vicinity of UV-chromophores.

In this publication we are reporting on selection of SAMPL6 pKa challenge compounds, their UV-metric pKa values measured by UV-metric titrations using Sirius T3, and NMR based microstate characterization of two SAMPL6 compounds (SM07 and SM14). We discuss implications of experimental technique for the interpretation of pKa data and suggestions for future pKa data collection efforts with the goal of evaluating or training computational pKa predictions.

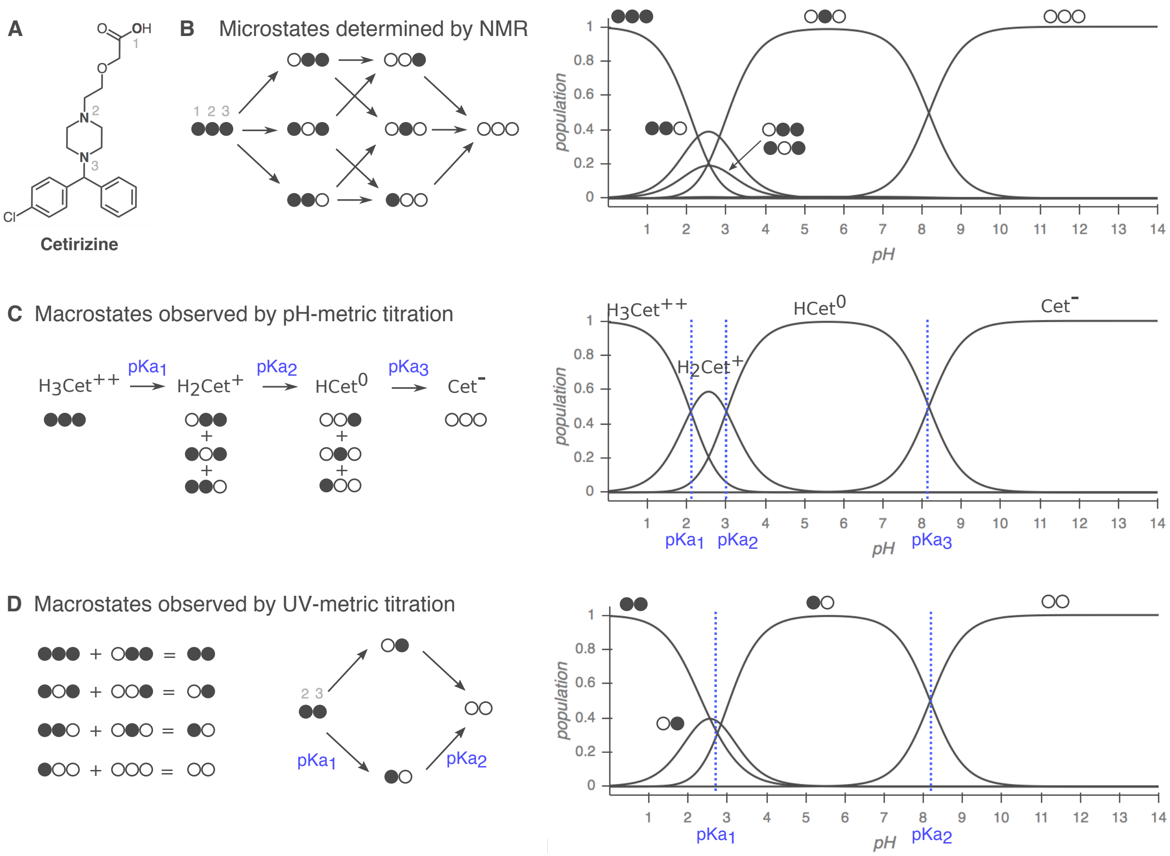


Figure 1. Comparison of macroscopic and microscopic pKa measurement methods. Illustration style of microstates was adopted from [5] and microscopic pKa values of cetirizine were measured by NMR [6]. Filled circles represent protonated sites and empty circles represent deprotonated sites with the order of carboxylic acid (1), piperazine nitrogen (2), and piperazine nitrogen (3). pProtonation state populations shown for pH-metric and UV-metric pKa measurement methods are simulations, calculated using NMR-based microscopic pKa values. **A** Cetirizine has 3 (n) titratable sites, shown in bold. **B** Left figure shows 8 microstates (2^n) and 12 micro-pKas ($n2^{n-1}$) of cetirizine. Right plot show relative population of microspecies with respect to pH. All microstates can be resolved with NMR method. **C** Simulated pH-metric(potentiometric) titration and macroscopic populations. For a polyprotic molecule, only macroscoopic pKas of can be measured with pH-metric titration. Microstates that have different number of protons bound can be resolved, but microstates that have the same total charge are observed as one macroscopic population. **D** Simulated microscopic populations for UV-metric (spectrophotometric) titration of cetirizine. Only protonation of the titration sites closer than 4 heavy atoms to UV-chromophore can cause change in UV-absorbance spectra, thus microstates that only differ by protonation of carboxylic acid can not be differentiated. Moreover, populations that overlap may or may not be resolved based on how different their absorbance spectra in UV-region are. Both UV-metric and pH-metric pKa determination methods measure macroscopic pKas for polyprotic molecules, which can not be easily assigned to individual titration sites and underlying microstate populations in the absence of other experimental evidence that provides structural information, such as NMR. Notice that macroscopic populations observed in these two methods are composed of different combinations of microstates depending on the principles of measurement technique.

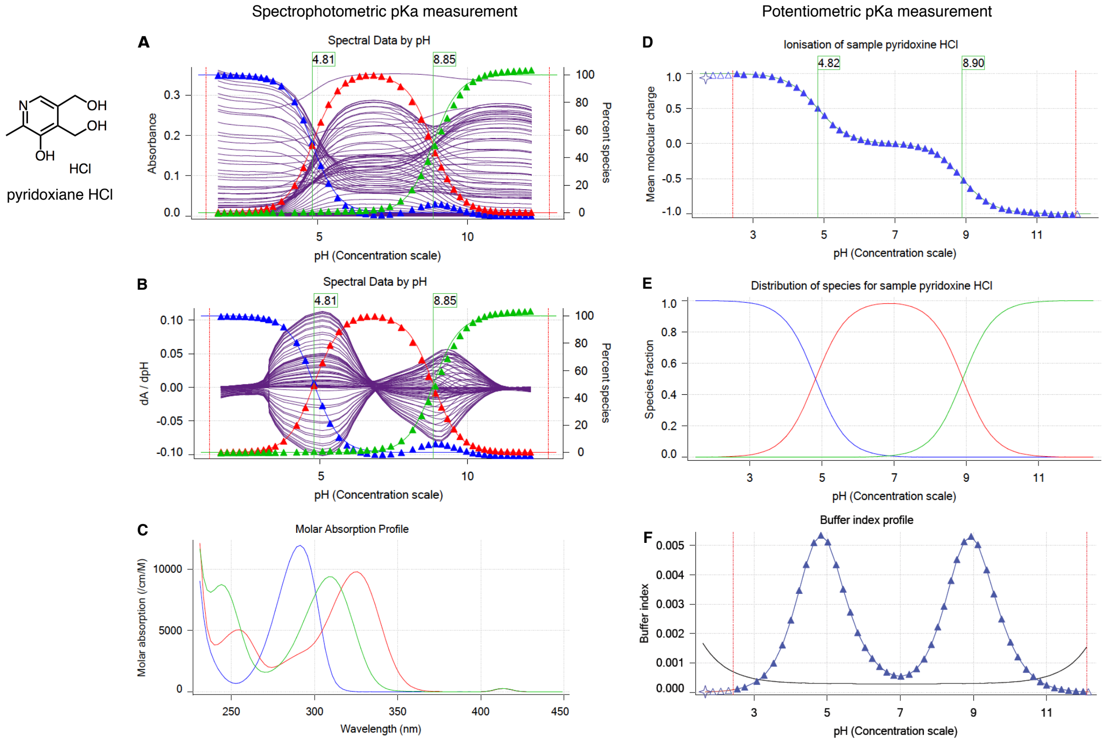


Figure 2. Spectrophotometric (UV-metric) and potentiometric (pH-metric) pKa measurements of pyridoxine HCl with Sirius T3. Spectrophotometic method (panels **A, B, C**) relies on UV-absorbance difference between protonation states. It is very sensitive and requires very low analyte concentrations(~ 50 μM), but limited to titration sites near chromophores. **A** Multiwavelength absorbance vs pH. Purple lines represents absorbance at each wavelength in UV-region. **B** Derivative of multiwavelength absorbance with respect to pH (dA/dpH) vs pH is plotted with purple lines. In **A** and **B** Blue, red, and green triangles represent population of protonation states (from most protonated to least protonated) as calculated from experimental data at each pH value. pKa values (green flags) correspond to inflection point of multiwavelength absorbance data where change in absorbance with respect to pH is maximum. **C** Molar absorption coefficients vs wavelength for each protonation state as resolved by TFA. **D, E, F** illustrate potentiometric pKa measurement where molar addition of acid or base is tracked as pH is titrated.**D** Mean molecular charge vs pH. Mean molecular charge is calculated based on the model provided for the analyte: predicted number and nature of titratable sites(acid or base type), and number of counter ions present. pKa values are calculated as inflection points of charge vs pH plot. **E** Predicted macroscopic protonation state populations vs pH calculated based on pKa values (H_2A^+: blue, HA: red, and A^-: green) **E** Buffering index vs pH profile of water (grey solid line, theoretical) and the sample solution (blue triangles represent experimental data points). A higher concentration of analyte (~ 5 mM) is necessary for potentiometric method than for spectrophotometric method to provide large enough buffering capacity signal above water for an accurate measurement.

Methods

Compound selection and procurement

To select a set of small molecules focusing on chemical space representing kinase inhibitors for physicochemical property prediction challenges (pKa and lipophilicity) we started from kinase targeting subclass of ZINC15 chemical library [8] and applied a series of filtering and selection rules as depicted in Figure 3 A. The following subsets of ZINC15 kinase subclass were selected in the first step: availability "now" and reactivity "anodyne", querried by: http://zinc15.docking.org/subclasses/kinase/substances/subsets/now+anodyne/. "Now" label indicates availability of compound for immediate delivery. "Anodyne" label excludes compounds matching reactivity patterns and pan-assay interference compounds (PAINs) [9, 10].

Next, we found which these molecules were also found in eMolecules [11] (free version of the database was downloaded on June 1st, 2017), since procurement was going to be done through eMolecules. To find the intersection of ZINC15 kinase subset and eMolecules database, we matched molecules using canonical isomeric SMILES as identifier. Canonical isomeric SMILES were created with OpenEye OEChem Toolkit [12]. About 100 mg of each compound in powder form and 90% purity was required complete planned experiments. To extract availability and price information from eMolecules, we queried using list of SMILES (as reported in eMolecules database) of the intersection set. We further filtered the intersection set of 1204 compounds based on delivery time (Tier 1 suppliers, 2 week delivery), at least 100 mg availability and in as powder (format: Supplier Standard Vial).

pKa measurements and logP measurements with Sirius T3 requires a titratable group in the pKa range of 2-12, so we wanted to select compounds with predicted pKas in the range of 3-11 considering possible errors in pKa predictions. pKa predictions were calculated using Epik Sequential pKa prediction (scan) [13, 14] with target pH 7.0 and tautomerization allowed for generated states. We filtered out all compounds that did not have any pKas predicted between 3-11 and also compounds with predicted pKa values closer than 1 pKa unit so that individual pKas of multiprotic compounds could be resolved with spectrophotometric pKa measurements. With the goal of selecting compounds suitable for logP measurements, we eliminated compounds with OpenEye XlogP [15] values lower than -1 and higher than 6. Subsets of compounds with molecular weights between 150-350 g/mol and 350-500 g/mol were selected for fragment-like and drug-like categories respectively. We eliminated compounds which didn't have price and stock amounts information available. Also compounds with publicly available experimental logP measurements were removed. The sources we checked for experimental logP values were the following: DrugBank [16] (queried with eMolecules SMILES), ChemSpider [17] (queried by canonical isomeric SMILES), NCI Open Database August 2006 release [18] and Enhanced NCI Database Browser [19] (queried with canonical isomeric SMILES), and PubChem [20] (queried with InChIKeys generated from canonical isomeric SMILES with NCI CACTUS Chemical Identifier Resolver [21].)

In order to include common ring structures found in kinase inhibitors, we analyzed the frequency of rings found in FDA-approved kinase inhibitors via Bemis-Murcko fragmentation [22, 23]. Heterocycles found more than once in FDA-approved kinase inhibitors are shown in Figure 3 B. While selection 25 compounds for fragment-like set and 10 compounds for drug-like set, We prioritized including at least one example of each heterocycle, although we failed to find compounds with piperazine and indazole that satisfied all other selection criteria. We observed that certain heterocycles (shown in Figure 3 C) were overrepresented based on our selection criteria, thus limited the number of these structures in SAMPL6 challenge set to at most one in fragment-like and drug-like sets. To achieve broad and uniform sampling of logP dynamic range, we segregated the molecules into bins of logP values and selected compounds from each bin prioritizing cheaper chemicals.

Presence of UV chromophores (absorbing in 200-400 nm) in proximity to protonation sites is necessary for spectrophotometric pKa measurements. To check for UV chromophores, we looked at the substructure matches to SMARTS pattern $[n, o, c][c, n, o]cc$ which were considered as the smallest unit of pi-conjugation that can constitute an UV-chromophore. This SMARTS pattern describes 4 heavy atom extended conjugation systems composed of aromatic carbon, nitrogen, or oxygen, such as 1.3-butadiene which absorbs at 217 nm. Additionally, final set of selected molecules were inspected to makes sure all potentially titratable groups were within 4 heavy atoms distance from a UV-chromophore. 25 fragment-like and 10 drug-like compounds were selected, out of which procurement and pKa measurements for 17(SM01-SM17) and 7(SM18-SM24) were successful, respectively. These 24 small molecules constituded the SAMPL6 pKa challenge set.

Python scripts used for compound selection are available at GitHub repository https://github.com/choderalab/sampl6-physicochemical-properties under **compound_selection/ zinc15_eMolecules_intersection_set** directory. Procurement details if each compound can be found in Table SI 1.

UV-metric pKa measurements

Experimental pKa measurements were collected using spectrophotometric pKa method with a Sirius T3 automated titrator instrument (Pion) at 25°C and constant ionic strength. Sirius T3 instrument is equiped with an Ag/AgCl double-junction reference electrode to monitor pH, a dip probe attached to spectrophotometer that measures, a stirrer, and automated volumetric titration capability. The UV-metric pKa measurement protocol of the Sirius T3 measures the change in multiwavelength absorbance in the 250-450 nm UV region of the absorbance spectrum while the pH is titrated between pH 1.8 and 12.2 to evaluate pKas [24, 25].

DMSO stock solutions of each compound with 10 mg/ml concentration were prepared by weighing 1 mg of powder chemical with Sartorius Analytical Balance (Model: ME235P) and dissolving it in 100 µL DMSO (Dimethyl sulfoxide, Fisher Bioreagents, CAT: BP231-100, LOT: 116070, purity ≥ 99.7%). DMSO stock solutions were capped immediately to limit hygroscopicity of DMSO and sonicated for 5-10 minutes in water bath sonicator at room temperature to ensure proper dissolution. These DMSO stock solutions were stored in room temperature up to 2 weeks. 10 mg/ml DMSO solutions were used as stock solutions for the preparation of 3 replicate samples for independent titrations: 15 µL of 10 mg/ml DMSO stock solution delivered to 4 mL glass sample vials of Sirius T3 with an electronic micropipette (Rainin EDP3 LTS 1-10 µL). The volume of delivered DMSO stock solution, which determines the sample concentration, is optimized individually for each compound to achieve sufficient but not saturated absorbance signal (targeting 0.5-1.0 AU) in the linear response region. Another limiting factor for sample concentration was ensuring that the compounds stays soluble in the whole range of pH titration range. 25 µL of mid-range buffer (14.7 mM K_2HPO_4 0.15 M KCl in H_2O) was added to each sample transfered with a micropipette (Rainin EDP3 LTS 10-100 µL) to provide enough buffering capacity in middle pH ranges so that pH could be controlled incrementally throughout the titration.

pH is temperature and ionic–strength dependent. Sample heating block of Sirius T3 kept the analyte solution at 25 ± 0.5 °C throughout the titration. Ionic–strength of the samples were adjusted by dilution in 1.5 mL ionic–strength adjusted water (ISA water, 0.15 M KCl) to the vials by Sirius T3. Analyte dilution, mixing, acid/base titration, and measurement of UV-absorbance was automated by UV-metric pKa measurement protocol of Sirius T3(Pion). pH was titrated between pH 1.8 and 12.2 with addition of acid (0.5 M HCl) and base (0.5 M KOH) targeting 0.2 pH steps between datapoints. Titrations were performed under Argon flow on the surface of sample solution to limit carbondioxide absorption from air. To capture all sources of experimental variability fully, instead of doing 3 tandem pH titrations on the same sample solution, pKas of three replicate samples were measured separately with one round of pH titration. Although this choice has higher cost for throughput and analyte consumption, it limits the dilution of the analyte during multiple titrations and drop in absorbance signal.

Visual inspection of sample solutions after titration and inspection of pH-dependent absorbance shift in 500-600 nm region of UV spectra was used to verify no detectable precipitation occurred during the course of measurement. Absorbance in 500-600 nm region of UV spectra is associated with scattering of longer wavelengths of light in the presence of aggregates. For each analyte, we optimized analyte concentration, direction of titration and the range of pH titration in order to achieve solubility. The direction of titration was determined so that titration would start from the pH where the compound is most soluble: low-to-high pH for bases and high-to-low pH for acids. UV-metric pKa measurement method typically requires as low as 50 µM sample concentration (although minimum concentration requirement may change based on absorbance properties of the analyte). Some compounds may not be solube even at such low concentration throughout pH range of the titration. As the sample is titrated through a wide range of pH values, it is likely that low solubility ionization states such as neutral and zwitterionic states will be also be visited, limiting the highest

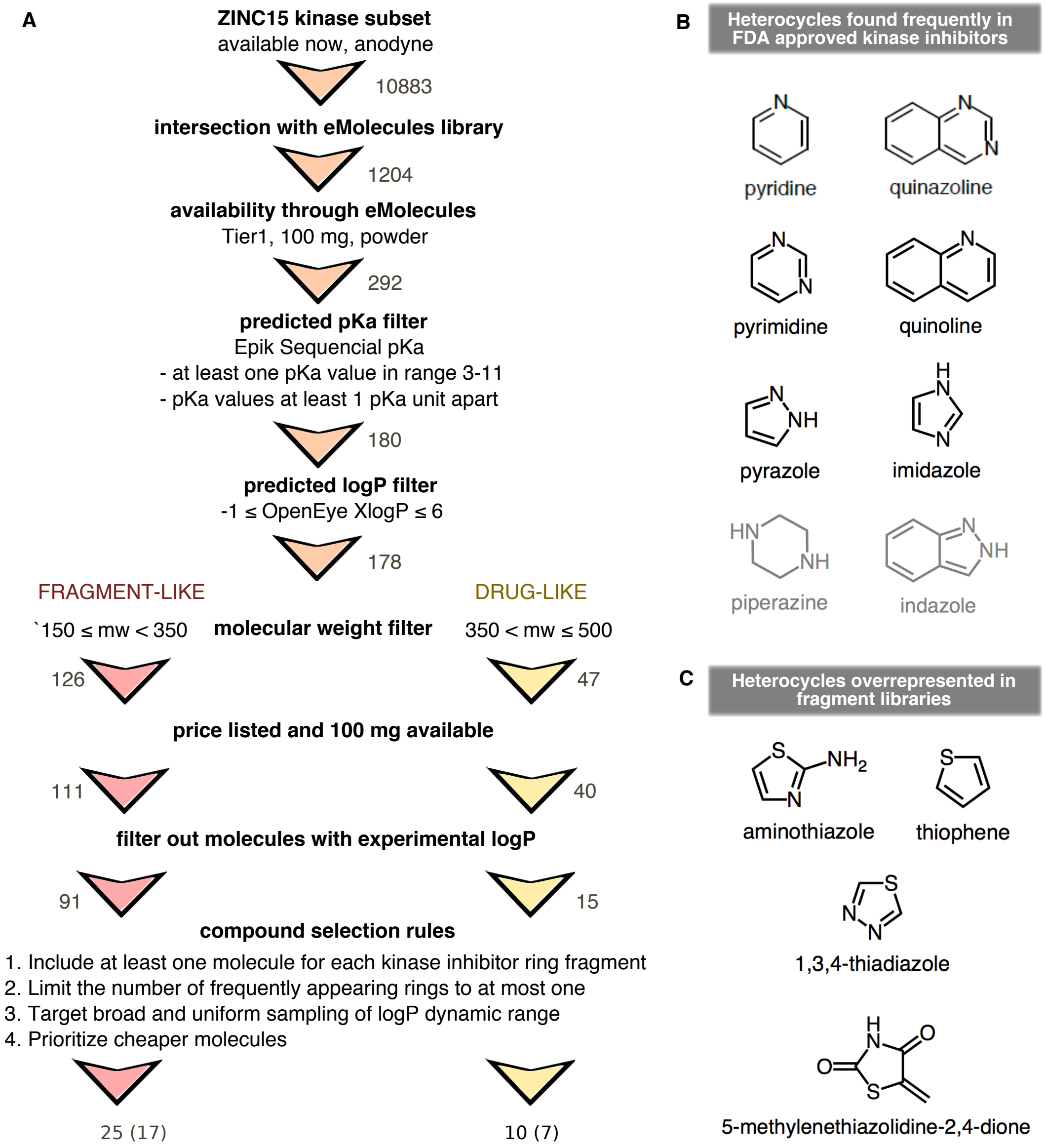

A
ZINC15 kinase subset
available now, anodyne
10883
intersection with eMolecules library
1204
availability through eMolecules
Tier1, 100 mg, powder
292
predicted pKa filter
Epik Sequencial pKa
- at least one pKa value in range 3-11
- pKa values at least 1 pKa unit apart
180
predicted logP filter
-1 ≤ OpenEye XlogP ≤ 6
178
FRAGMENT-LIKE
DRUG-LIKE
`150 ≤ mw < 350
molecular weight filter
350 < mw ≤ 500
126
47
price listed and 100 mg available
111
40
filter out molecules with experimental logP
91
15
compound selection rules
1. Include at least one molecule for each kinase inhibitor ring fragment
2. Limit the number of frequently appearing rings to at most one
3. Target broad and uniform sampling of logP dynamic range
4. Prioritize cheaper molecules
25 (17)
10 (7)
B
Heterocycles found frequently in FDA approved kinase inhibitors
pyridine
quinazoline
pyrimidine
quinoline
pyrazole
imidazole
piperazine
indazole
C
Heterocycles overrepresented in fragment libraries
aminothiazole
thiophene
1,3,4-thiadiazole
5-methylenethiazolidine-2,4-dione

Figure 3. Compound selection for SAMPL6 physicochemical properties challenges for pKa and logP. A Flowchart of filtering steps for the selection of SAMPL6 compounds similar to kinase inhibitors and kinase inhibitor fragments. Numbers next to arrows how number of potential compounds after each filtering step. 25 fragment-like and 10 drug-like compounds were selected, out of which procurement and pKa measurements for 17 and 7 were successful, respectively. **B** Frequent heterocycles found in FDA approved kinase inhibitors determined by Bemis-Murcko fragmentation into rings [22]. Black structures were represented in SAMPL6 set at least once. Compounds with piperazine and indazole (gray structures) could not be included in the challenge set due to library and selection limitations. **C** Structures of heterocycles that were overrepresented based on our compound selection workflow. We have limited occurance of these heterocycles to at most one.

analyte concentration that can be titrated without encountering any solubility issues. For compounds with insufficient solubility to accurately determine a pKa directly in a UV-metric titration, a cosolvent protocol was used [See the next section: UV-metric pKa measurement with cosolvent].

Two Sirius T3 softwares, Sirius T3Control v1.1.3.0 and Sirius T3Refine v1.1.3.0, were used to execute measurement protocols and to analyze pH-dependent multivelenghth spectra, respectively. A protonation state change of titratable sites near chromophores will modulate the UV-absorbance spectra of these chromophores, allowing populations of distinct UV-active species to be resolved as a function of pH. To do this, basis spectra are identified and populations extracted via TFA analysis of the pH-dependent multi-wavelength absorbance [25]. When fitting the absorbance data to a titratable molecule model to estimate pKas, we selected the minimum number of pKas that was sufficient to provide a high quality of fit between experimental and modeled data based on visual inspection of pH-dependent populations.

This method is capable of measuring pKas between 2 and 12 when protonatable groups are at most 4-5 heavy atoms away from chromophores such that a change in protonation state alters the absorbance spectrum of the chromophore. We have selected compounds where titratable groups are close to potential chromophores (generally aromatic ring systems), but it is possible that our experimental results couldn't detect protonation of titratable groups distal to a UV-chromophore.

UV-metric pKa measurement with cosolvent

If analytes are not sufficiently soluble in aqueous environment, pKa values can not be determined accurately with UV-metric pKa measurements method. If precipitation is present, the UV-absobance signal from pH-dependent precipitate formation can not be differentiated from the pH-dependent protonation signal. For compounds with low aqueous solubility pKa values were estimated from multiple apparent pKas measurements done in methanol:water cosolvent solutions with various ratios. This method is refered to as UV-metric psKa method in Sirius T3 Manual [26].

In cosolvent spectrophotometric pKa measurement protocol was very similar to UV-metric pKa measurement method in water except the following aspects: Three titrations were performed in typically in 30%, 40%, and 50% mixtures methanol:water by volume to measure apparent pKa values (psKa) in these mixtures. Than Yasuda-Shedlovsky extrapolation method was used to estimate the pKa value at 0% cosolvent [27–29].

$$psKa + log[\mathrm{H_2O}] = A/\epsilon + B$$

Yasuda-Shedlovsky extrapolation relies on the linear correlation between $psKa + log[\mathrm{H_2O}]$ and reciprocal dielectric constant of the cosolvent mixture ($1/\epsilon$). In the equation above, A and B are the slope and intercept of the line fitted to experimental datapoints. Depending on the solubility requirements of the analyte methanol ratio of the cosolvent mixtures were adjusted. We designed the experiments to have at least 5% cosolvent ratio difference between datapoints and 60% methanol at most. Calculation of Yasuda-Shedlovsky extrapolation was performed by Sirius T3 software using at least 3 psKa values measured in different ratios of methanol:water. Addition of methanol (ISA, 0.15 M KCl) was controlled by the instrument before each titration. Three consecutive pH titrations at different methanol concentrations were performed using the same sample solution. In addition, three replicate measurements with independent samples (prepared from the same DMSO stock) were collected.

Calculation of uncertainty in pKa measurements

Experimental uncertainties were reported as standard error of the mean (SEM) of three replicate pKa measurements. Since Sirius T3 software reports pKa values with 2 decimal places, we have reported SEM as 0.01 in cases where SEM values calculated from 3 replicates were lower than 0.01.

Protonation site determination with NMR

Iyke will provide the text for NMR method.

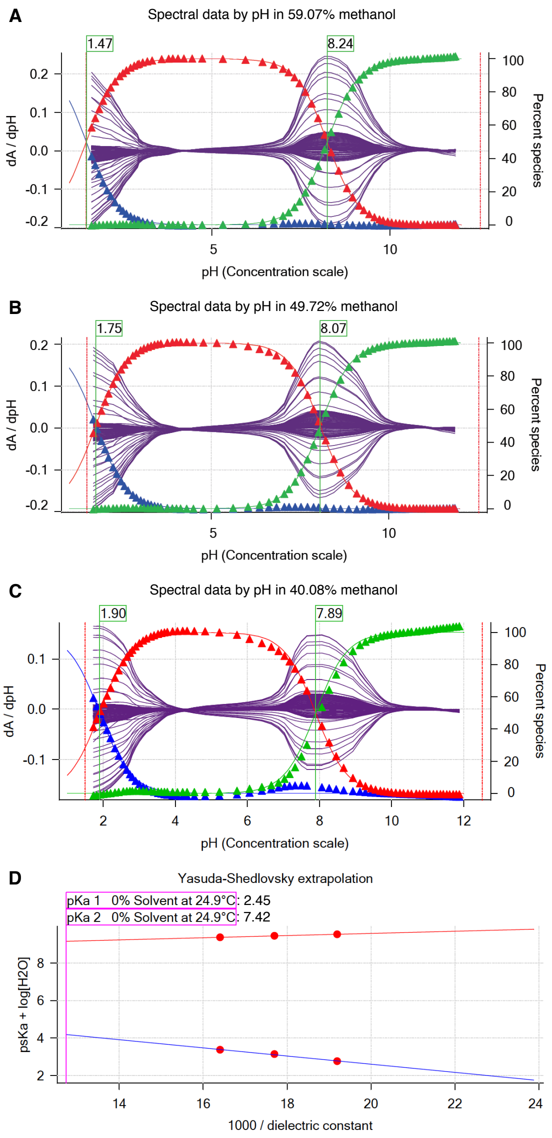


Figure 4. Determination of SM22 pKa values with cosolvent method and Yasuda-Shedlovsky extrapolation. A, **B**, and **C** show psKa of SM22 determined at various methanol concentrations: 59.07%, 49.72%, 40.08% by weight. Purple solid lines indicate derivative of absorbance with respect to pH vs pH in multiple wavelength. psKa values (green flags) were determined by Sirius T3 Refinement Software. Blue, red and green triangles show relative populations of macroscopic protonation states with respect to pH calculated from the experimental data. Notice that as cosolvent concentration increases, psKa1 value decreases from 1.90 to 1.47 and psKa2 value increases from 7.84 to 8.24. **D** Yasuda-Shedlovsky extrapolation plot for SM22. Red datapoints correspond to psKa determined at various cosolvent ratios. Based on linear fitting to $psKa + log[\mathrm{H_2O}]$ vs $1//\epsilon$, pKa1 and pKa2 in 0% cosolvent (aqueous solution) was determined as 2.45 and 7.42, respectively. R^2 values of linear fits are both 0.99. The slope of Yasuda-Shedlovsky extrapolation shows if the observed titration has acidic(positive slope) or basic(negative) character dominantly, although this is an macroscopic observation and should not be relied on for annotation of pKas to functional groups (microscopic pKas).

Quality control for chemicals

Purity of SAMPL6 pKa challenge compounds were determined based on LC–MS. The purity analysis was performed using an Agilent HPLC1200 Series equipped with auto–sampler, UV diode array detector and a Quadrupole MS detector 6140. The software used was Chemistation for LC & LCMS with version C01.07SR2. The column for the analysis is Assentis Express C18 3.0x100mm 2.7 µl particle size at Column temperature = 45°C.

- Mobile phase A: 2 mM ammonium formate (pH=3.5) aqueous
- Mobile phase B: 2 mM ammonium formate in Acetonitrile : Water=90:10 (pH=3.5)
- Flow rate : 0.75ml/min
- Gradient: Starting with 10% B to 95%B in 10 minutes then hold 95%B for 5 minutes.
- Post run length: 5 minutes
- Mass condition: ESI positive and negative mode
- Capillary voltage: 3000 V
- Grying gas flow: 12 ml/min
- Nebulizer pressure: 35 psi
- Drying temperature: 350°C
- Mass range: 5-1350 Da; Fragmentor:70; Threshold:100

The percent area for main peak is calculated based on the area of main peak divided by the total area of all peaks. The percent area of the main peak is reported as an estimate of sample purity.

Results

Spectrophotometric pKa measurements

todo[inline]Summarize measurement results, refer to tables and SI.

Impact of cosolvent to UV-metric pKa measurements

For molecules with insufficient solubility in aqueous medium throughout the titration range (pH 2-12), we had to use UV-metric pKa measurement with cosolvent methanol. To confirm that UV-metric pKa measurements with cosolvent method lead to indistinguishable results compared to UV-metric measurements in water, we collected pKa values of 12 highly soluble SAMPL6 compounds and pyridoxine also with cosolvent method. Correlation analysis of pKa values measured with two methods show that using methanol as cosolvent and determining aqueous pKas via Yasuda-Shedlovsky extrapolation did not cause any bias (Figure 6).

Purity of SAMPL6 compounds

LC-MS based purity measurements showed that powder stocks of 23 SAMPL6 pKa challenge compounds were >90% pure and purity of SM22 was 87% which was the lowest in the set (Table SI 6). Additionally, molecular weights detected by LC-MS method were consistent with molecular weights reported in eMolecules and supplier reported molecular weights when provided.

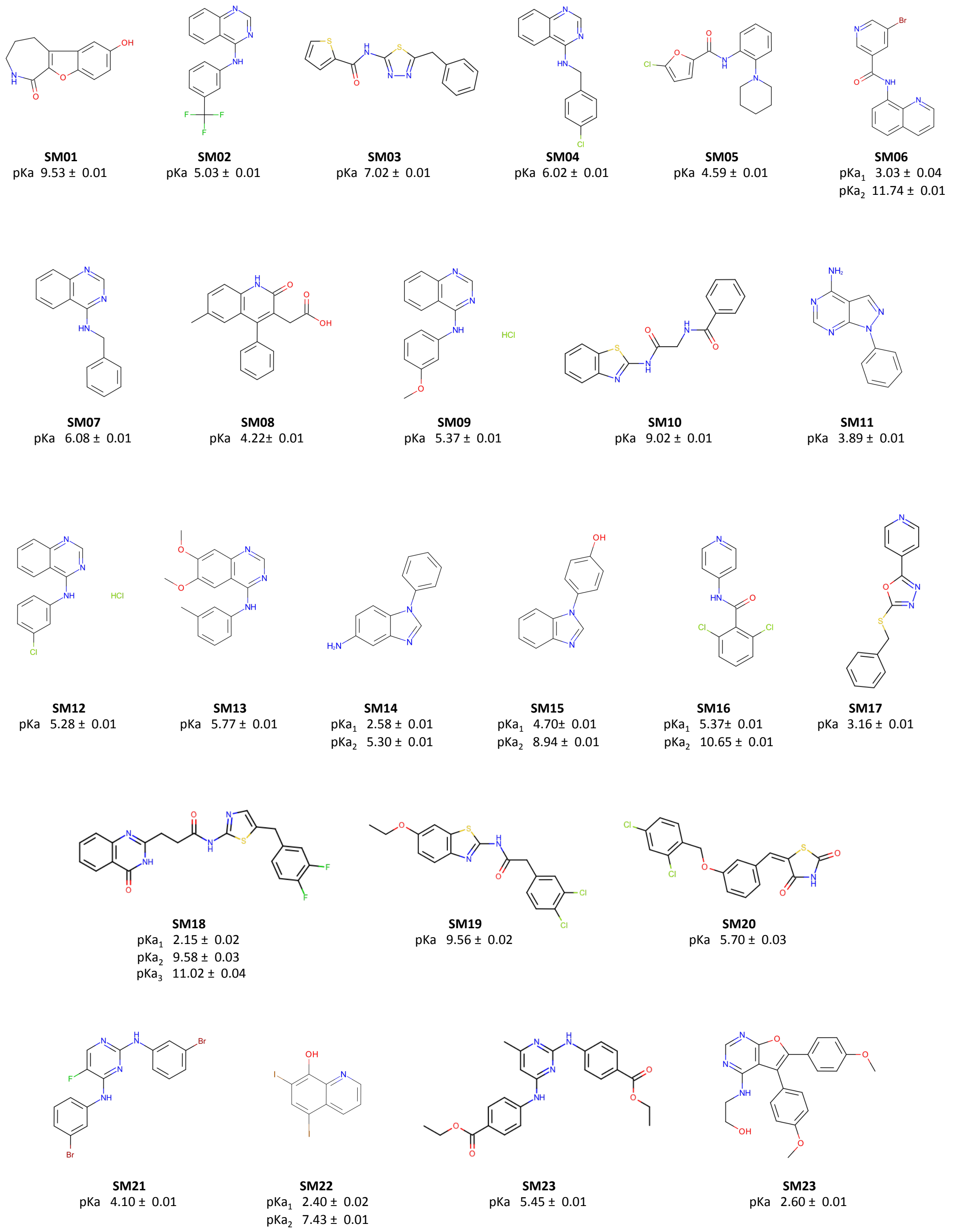
Characterization of SM07 microstates with NMR

Add results of NMR analysis of SM07 - from Iyke

Discussion

Discussion of experimental data interpretation: macroscopic

Multiwavelength absorbance analysis on thw Sirius T3 allows for very good resolution of pKas, but it is important to note that this method produces estimates of macroscopic pKas. If multiple microscopic pKas have close pKa values and overlapping changes in UV absorbance spectra associated with protonation/deprotonaton event, the spectral analysis could produce a single macroscopic pKa that represents an aggregation of multiple microscopic pKas.



SM01
pKa 9.53 ± 0.01

SM02
pKa 5.03 ± 0.01

SM03
pKa 7.02 ± 0.01

SM04
pKa 6.02 ± 0.01

SM05
pKa 4.59 ± 0.01

SM06
pKa_1 3.03 ± 0.04
pKa_2 11.74 ± 0.01

SM07
pKa 6.08 ± 0.01

SM08
pKa 4.22± 0.01

SM09
pKa 5.37 ± 0.01

SM10
pKa 9.02 ± 0.01

SM11
pKa 3.89 ± 0.01

SM12
pKa 5.28 ± 0.01

SM13
pKa 5.77 ± 0.01

SM14
pKa_1 2.58 ± 0.01
pKa_2 5.30 ± 0.01

SM15
pKa_1 4.70± 0.01
pKa_2 8.94 ± 0.01

SM16
pKa_1 5.37± 0.01
pKa_2 10.65 ± 0.01

SM17
pKa 3.16 ± 0.01

SM18
pKa_1 2.15 ± 0.02
pKa_2 9.58 ± 0.03
pKa_3 11.02 ± 0.04

SM19
pKa 9.56 ± 0.02

SM20
pKa 5.70 ± 0.03

SM21
pKa 4.10 ± 0.01

SM22
pKa_1 2.40 ± 0.02
pKa_2 7.43 ± 0.01

SM23
pKa 5.45 ± 0.01

SM23
pKa 2.60 ± 0.01

Figure 5. Molecules used in the SAMPL6 pKachallenge. Experimental UV-metric pKameasurements were performed for these 24 molecules, and discernable macroscopic pKas reported. Uncertainties are expressed as the standard error of the mean (SEM) of three independent measurements. Structure figures were created with OpenEye OEDepict Toolkit [30]

Table 1. Experimental pKas of SAMPL6 compounds. Spectrophotometric pKa measurements were performed with two assay types based on aqueous solubility of analytes. "UV-metric pKa" assay indicates spectrophotometric pKa measurements done with Sirius T3 in ISA water. "UV-metric pKa with cosolvent" assay refers to pKa determination by Yasuda-Shedlovsky extrapolation from psKa measurements in various ratios of ISA methanol:water mixtures. Triplicate measurements were performed at 25 ± 0.5 °C and in the presence of approximately 150 mM KCl to adjust ionic strength.

Molecule ID	pKa_1	pKa_2	pKa_3	Assay Type
SM01	9.53 ± 0.01			UV-metric pKa
SM02	5.03 ± 0.01			UV-metric pKa with cosolvent
SM03	7.02 ± 0.01			UV-metric pKa with cosolvent
SM04	6.02 ± 0.01			UV-metric pKa
SM05	4.59 ± 0.01			UV-metric pKa with cosolvent
SM06	3.03 ± 0.04	11.74 ± 0.01		UV-metric pKa
SM07	6.08 ± 0.01			UV-metric pKa
SM08	4.22 ± 0.01			UV-metric pKa
SM09	5.37 ± 0.01			UV-metric pKa with cosolvent
SM10	9.02 ± 0.01			UV-metric pKa with cosolvent
SM11	3.89 ± 0.01			UV-metric pKa
SM12	5.28 ± 0.01			UV-metric pKa
SM13	5.77 ± 0.01			UV-metric pKa
SM14	2.58 ± 0.01	5.30 ± 0.01		UV-metric pKa
SM15	4.70 ± 0.01	8.94 ± 0.01		UV-metric pKa
SM16	5.37 ± 0.01	10.65 ± 0.01		UV-metric pKa
SM17	3.16 ± 0.01			UV-metric pKa
SM18	2.15 ± 0.02	9.58 ± 0.03	11.02 ± 0.04	UV-metric pKa with cosolvent
SM19	9.56 ± 0.02			UV-metric pKa with cosolvent
SM20	5.70 ± 0.03			UV-metric pKa with cosolvent
SM21	4.10 ± 0.01			UV-metric pKa with cosolvent
SM22	2.40 ± 0.02	7.43 ± 0.01		UV-metric pKa with cosolvent
SM23	5.45 ± 0.01			UV-metric pKa with cosolvent
SM24	2.60 ± 0.01			UV-metric pKa with cosolvent

[1] pKa values are reported as mean and SEM of three replicates.

Warning about acid base assingment from cosolvent measurements

Lessons learned for future challenge iterations

Discussion of NMR results

The goal of this NMR characterization was collecting information on microscopic states related to experimental pKa measurements, i.e. determining sites of protonation. pKa measurements performed with spectrophotometric method provides macroscopic pKa values, but does not provide information site of protonation. On the other hand, most computational prediction methods predict primarily microscopic pKa values. Protonation sites can be determined by NMR methods, although these measurements are very laborious in terms of data collection and interpretation compared to pKa measurements with Sirius T3. Moreover, not all SAMPL6 molecules were suitable for NMR measurements due to high sample concentration requirements (for methods other than proton NMR) and analyte solubility issues. Thus we performed NMR based microstate characterization only for SM07. We investigated microstates existed at pH values lower and higher than macroscopic pKa value with the goal of evaluating if spectroscopicly measured pKa was microscopic (related to single protonation site).

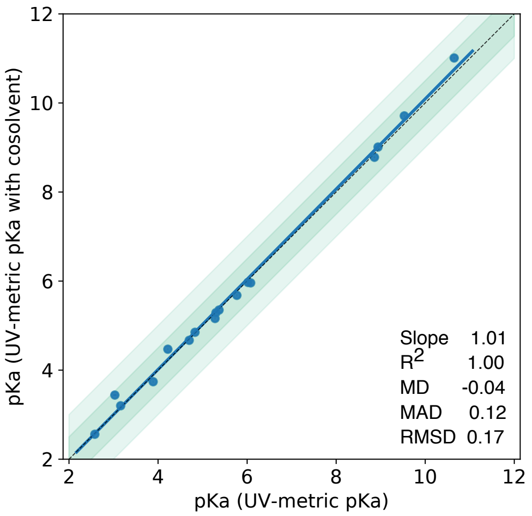


Figure 6. pKa measurements with UV-metric method with cosolvent and UV-metric method in water show good correlation. 17 pKa values(blue marks) of 13 chemicals were measured with both UV-metric pKa method in water and UV-metric pKa method with methanol as cosolvent (Yasuda-Shedlovsky extrapolation to 0% methanol). Dashed black line has slope of 1, representing perfect correlation. Dark and light green shaded areas indicate +-0.5 and +-1.0 pKa unit difference regions, respectively. Error bars are plotted as SEM of replicate measurements, although they are not visible since the largest SEM value is 0.04. MD: Mean difference, MAD: Mean absolute deviation, RMSD: Root-mean-square deviation.

Interpretation of NMR microstates for 4-amino quinazoline series

In addition to SM07, there were 5 other 4-amino quinazoline derivatives in SAMPL6 set: SM02, SM04. SM09, SM12, and SM13. Based on structural similarity, we can infer that spectrometric pKa values measured for other 4-amino quinazoline compounds as microscopic pKa related to the protonation of the same quinazoline nitrogen with the same neutral background protonation states.

Challenges with NMR measurments?

pKa can shift based on ionic strength, temperature, and lipophilic content Can other cosolvents be interesting for process chemistry.

Code and data availability

- SAMPL6 pKa challenge instructions, submissions, and analysis is available at https://github.com/MobleyLab/SAMPL6
- Python scripts used for compound selection are available at **compound_selection** directory of https://github.com/choderalab/sampl6-physicochemical-properties

Construct a proper README for compound selection directory.

Overview of supplementary information

Organized in SI document:

- TABLE SI 1: procurement details of SAMPL6 compounds
- TABLE SI 2: selection details of SAMPL6 compounds
- TABLE SI 3: pKa results of replicate experiments csv
- TABLE SI 4: pKa results of water and cosolvent replicate experiments csv
- TABLE SI 5: pKa mean and SEM results of water and cosolvent replicate experiments

- NMR spectra of SM07 microstate characterization
- TABLE SI 6: Summary of LC-MS purity results
- LC-MS Figures

Extra files:

- Sirius T3 reports

Author Contributions

Complete this section.

Conceptualization, ; Methodology, ; Software, ; Formal Analysis, ; Investigation, MI, ; Resources, ; Data Curation, ; Writing-Original Draft, ; Writing - Review and Editing, ; Visualization, ; Supervision, ; Project Administration, ; Funding Acquisition,

Acknowledgments

Complete this section.

MI, ASR, and JDC acknowledge support from the Sloan Kettering Institute. JDC acknowledges support from NIH grant P30 CA008748. MI acknowledges Doris J. Hutchinson Fellowship. Acknowledging Merck Preformulation department for materials, expertise and instrument time.

Brad Sherborne Paul Czodrowski - helped with purchasable compound list in earlier iteration Ikenna Ndukwe Caitlin Bannan

Disclosures

JDC is a member of the Scientific Advisory Board for Schrödinger, LLC.

References

[1] **Bannan CC**, Burley KH, Chiu M, Shirts MR, Gilson MK, Mobley DL. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. Journal of Computer-Aided Molecular Design. 2016 Nov; 30(11):927–944. http://link.springer.com/10.1007/s10822-016-9954-8, doi: 10.1007/s10822-016-9954-8.

[2] **Rustenburg AS**, Dancer J, Lin B, Feng JA, Ortwine DF, Mobley DL, Chodera JD. Measuring experimental cyclohexane-water distribution coefficients for the SAMPL5 challenge. Journal of Computer-Aided Molecular Design. 2016 Nov; 30(11):945–958. http://link.springer.com/10.1007/s10822-016-9971-7, doi: 10.1007/s10822-016-9971-7.

[3] **Pickard FC**, König G, Tofoleanu F, Lee J, Simmonett AC, Shao Y, Ponder JW, Brooks BR. Blind prediction of distribution in the SAMPL5 challenge with QM based protomer and pK a corrections. Journal of Computer-Aided Molecular Design. 2016 Nov; 30(11):1087–1100. http://link.springer.com/10.1007/s10822-016-9955-7, doi: 10.1007/s10822-016-9955-7.

[4] **Darvey IG**. The assignment of pKa values to functional groups in amino acids. Wiley Online Library; 1995.

[5] **Rupp M**, Korner R, V Tetko I. Predicting the pKa of small molecules. Combinatorial chemistry & high throughput screening. 2011; 14(5):307–327.

[6] **Marosi A**, Kovács Z, Béni S, Kökösi J, Noszál B. Triprotic acid–base microequilibria and pharmacokinetic sequelae of cetirizine. European Journal of Pharmaceutical Sciences. 2009 Jun; 37(3-4):321–328. http://linkinghub.elsevier.com/retrieve/pii/S0928098709000773, doi: 10.1016/j.ejps.2009.03.001.

[7] **Comer JEA**, Manallack D. Ionization Constants and Ionization Profiles. In: *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* Elsevier; 2014.http://linkinghub.elsevier.com/retrieve/pii/B9780124095472112338, doi: 10.1016/B978-0-12-409547-2.11233-8.

[8] **Sterling T**, Irwin JJ. ZINC 15 – Ligand Discovery for Everyone. Journal of Chemical Information and Modeling. 2015 Nov; 55(11):2324–2337. http://pubs.acs.org/doi/10.1021/acs.jcim.5b00559, doi: 10.1021/acs.jcim.5b00559.

[9] **Baell JB**, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. Journal of Medicinal Chemistry. 2010 Apr; 53(7):2719–2740. http://pubs.acs.org/doi/abs/10.1021/jm901137j, doi: 10.1021/jm901137j.

[10] Saubern S, Guha R, Baell JB. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. Molecular Informatics. 2011 Oct; 30(10):847–850. http://doi.wiley.com/10.1002/minf.201100076, doi: 10.1002/minf.201100076.

[11] eMolecules Database Free Version;. Accessed: 2017-06-01. https://www.emolecules.com/info/products-data-downloads.html.

[12] OEChem Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[13] Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. Journal of Computer-Aided Molecular Design. 2007 Dec; 21(12):681–691. http://link.springer.com/10.1007/s10822-007-9133-z, doi: 10.1007/s10822-007-9133-z.

[14] Schrödinger Release 2016-4: Epik Version 3.8;. Schrödinger, LLC, New York, NY, 2016.

[15] OEMolProp Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[16] Wishart DS. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research. 2006 Jan; 34(90001):D668–D672. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkj067, doi: 10.1093/nar/gkj067.

[17] Pence HE, Williams A. ChemSpider: An Online Chemical Information Resource. Journal of Chemical Education. 2010 Nov; 87(11):1123–1124. http://pubs.acs.org/doi/abs/10.1021/ed100697w, doi: 10.1021/ed100697w.

[18] NCI Open Database, August 2006 Release;. https://cactus.nci.nih.gov/download/nci/.

[19] Enhanced NCI Database Browser 2.2;. https://cactus.nci.nih.gov/ncidb2.2/.

[20] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Research. 2016 Jan; 44(D1):D1202–D1213. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv951, doi: 10.1093/nar/gkv951.

[21] NCI/CADD Chemical Identifier Resolver;. https://cactus.nci.nih.gov/chemical/structure.

[22] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. Journal of medicinal chemistry. 1996; 39(15):2887–2893.

[23] OEMedChem Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[24] Tam KY, Takács-Novák K. Multi-wavelength spectrophotometric determination of acid dissociation constants: a validation study. Analytica chimica acta. 2001; 434(1):157–167.

[25] Allen RI, Box KJ, Comer JEA, Peake C, Tam KY. Multiwavelength spectrophotometric determination of acid dissociation constants of ionizable drugs. Journal of pharmaceutical and biomedical analysis. 1998; 17(4):699–712.

[26] Sirius T3 User Manual, v1.1. Sirius Analytical Instruments Ltd, East Sussex, UK; 2008.

[27] Avdeef A, Box KJ, Comer JEA, Gilges M, Hadley M, Hibbert C, Patterson W, Tam KY. PH-metric logP 11. pK a determination of water-insoluble drugs in organic solvent–water mixtures. Journal of pharmaceutical and biomedical analysis. 1999; 20(4):631–641.

[28] Avdeef A, Comer JEA, Thomson SJ. pH-Metric log P. 3. Glass electrode calibration in methanol-water, applied to pKa determination of water-insoluble substances. Analytical Chemistry. 1993; 65(1):42–49. https://doi.org/10.1021/ac00049a010, doi: 10.1021/ac00049a010.

[29] Takács-Novák K, Box KJ, Avdeef A. Potentiometric pKa determination of water-insoluble compounds: validation study in methanol/water mixtures. International Journal of Pharmaceutics. 1997; 151(2):235 – 248. http://www.sciencedirect.com/science/article/pii/S0378517397049077, doi: https://doi.org/10.1016/S0378-5173(97)04907-7.

[30] OEDepict Toolkit 2017.Feb.1;. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

[31] Manallack DT. The pK(a) Distribution of Drugs: Application to Drug Discovery. Perspectives in Medicinal Chemistry. 2007 Sep; 1:25–38.